刘　雷,王　星. 精准医学知识库的构建[J]. 中华医学图书情报杂志,2018,27(6):1-9.

DOI:10.3969/j.issn.1671-3982.2018.06.001

·专题·

专家简介

刘　雷(1966-),男,北京市人,博士,教授,博士生导师。现任复旦大学生物医学研究院教授、复旦大学大数据研究院医学信息与医学影像智能诊断研究所所长、十三五国家精准医学重点发计划“疾病研究精准医学知识库构建”项目首席科学家、中国研究型医院学会临床数据与样本资源库专业委员会副主任委员、中国医药生物技术协会组织生物样本库分会常务委员。主要研究方向为生物医学大数据、精准医学以及医学人工智能,发表 SCI 论文百余篇。E-mail:liulei@fudan.edu.cn

精准医学知识库的构建

刘　雷,王　星

[摘要]根据“精准医学本体和语义表示标准”,构建了“精准医学知识库”及“精准医学知识库管理与共享平台”。该精准医学知识库体系具有标准规范、信息全面、开放共享、用户友好以及动态更新等特点,将全面支撑精准医学基础研究和临床应用,填补了国内精准医学知识库领域的空白。

[关键词]精准医学;知识库;精准医学本体;知识识别;知识推理;知识模型;知识图谱;文本挖掘

[中图分类号]R-058　　[文献标志码]A　　[文章编号]1671-3982(2018)06-0001-09

Development of knowledge base for precision medicine

LIU Lei, WANG Xing

(Fudan University, Shanghai 200032, China)

[**Abstract**]A knowledge base for precision medicine was developed, its management and sharing platform was built according to the precision medicine ontology and semantic presentation standards, which is characterized by its normalized standards, all-inclusive information, open sharing, user friendly and dynamic update. The knowledge base for precision medicine we developed can be used in basic research and clinical practice of precision medicine and has thus filled the gap in domestic knowledge base of precision medicine.

[**Key words**]Precision medicine; Knowledge base; Precision medicine ontology; Knowledge identification; Knowledge reasoning; Knowledge model; Knowledge mapping; Text mining

1　精准医学的发展

精准医学是生物技术、信息技术和多种前沿技术在医学临床实践的交汇融合应用,是医学科技发展的前沿方向,实施精准医学已经成为推动全民健康的国家发展战略。2011 年美国国家研究理事会提出“精准医学”概念,随着相关技术发展以及对该理念的重视,2015 年美国将“精准医学计划”提升为国家战略之一。精准医学的有效实施不仅可以提高国民健康和医疗水平,也可以更好地优化国家医疗资源分配,同时推动相关学科和技术的快速发展及相关产业的发展,进而产生巨大市场空间。因此,精准医学研究已经成为各国的科技战略制高点。2015 年,我国科技部召开了“国家精准医疗战略专家会

[基金项目]国家重点研发计划“精准医学研究”重点专项“疾病研究精准医学知识库构建”项目(2016YFC09011900)

[作者单位]复旦大学,上海　200032

议”,成立了中国精准医疗战略专家组,计划将在2030 年前投入 600 亿元,用于我国精准医学研究。2016 年 3 月,科技部官网公布了《科技部关于发布国家重点研发计划精准医学研究等重点专项 2016 年度项目申报指南的通知》。至此,精准医疗在中国已经上升为“国家战略”。复旦大学有幸成为精准医学重点专项中的首席单位。

精准医学研究集合现代医学和生物学科技发展的知识与技术,代表现代医学的发展趋势以及临床实践发展方向。精准医学的核心思想是通过对大样本、海量数据进行整合分析,构建能够揭示个体疾病分子机制的知识网络,由此针对病人的基因组和其他个体特点进行预防和治疗。随着生物医学领域研究的飞速发展,生物医学数据呈指数级增长,导致科研人员和医生难以从海量生物医学数据中发现高质量、可用性的知识。

自从人类基因组计划以来,测序技术和质谱技术等各类组学技术的飞速发展,推动了基因组、转录组、表观遗传组、蛋白质组和代谢组等海量生命科学组学数据的指数级增长[1-2]。一方面,机器学习和人工智能技术的发展大幅提升了医学信息学和生物信息学的发展,因此生物医学数据的应用方式也发生了改变。高通量实验技术的突破,直接把生物医学数据从 PB(PetaByte)时代推升到多维度数据融合的 EB(ExaByte)时代。另一方面,人群队列研究、分子流行病学研究产生了大量的数据,从分子、细胞、组织、器官、个体等多层面描述的多维度数据,汇总海量真实世界数据(Real World Data)[3-4],这些广泛的数据构成了复杂的高维度生物医学大数据。

目前生物医学数据具有数量巨大、增长迅速、质量控制困难、来源广泛繁杂、内涵丰富、非标准化、非结构化和数据相对分散等特点,从而导致难以挖掘生物医学大数据的潜在高价值。面对海量的生物医学数据,亟需构建精准医学知识库,全面获取各类生物医学文本信息和组学数据,在标准、统一的语义网络下,通过挖掘、关联等技术,从海量信息中高效准确地发现知识,为研究和临床决策提供充分可靠的依据,最终实现精准预防、精准诊断和精准治疗的目标。

因此,构建能够对海量数据进行分析并提供可靠知识的精准医学知识库,成为精准医学研究和临床应用发展的关键环节。

2 生物医学知识库现状

2.1 国外生物医学知识库建设

随着精准医学的发展,生物医学知识库成为生物医学领域研究的热点。美国国立生物医学中心开发了基于位点变异-基因-疾病的知识库 ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/)[5],欧洲生物信息研究所先后开发了蛋白质相互作用数据库 IntAct(https://www.ebi.ac.uk/intact/)[6]、生物学通路知识库 Reactome (https://reactome.org/)[7]、生物相关的化学实体数据库 ChEBI(https://www.ebi.ac.uk/inc/tool/chebi.html)、生化反应的数学模型数据库 BioModels (http://www.ebi.ac.uk/biomodels-main/) 和基因本体数据库 Gene Ontology (http://www.geneontology.org/) 等,西班牙国家生物技术中心开发了以基因为中心、基于 PubMed 文献摘要的在线文本知识挖掘服务平台 Information Hyperlinked over Proteins(iHOP),并将其用于提供基因间关联挖掘和分析。与此同时,一些公司也开展了生物医学知识库的开发,代表性平台有 GeneGo (https://portal.genego.com/)、IPA(http://www.ipa-world.org/) 和 Pathway Studio (http://www.pathwaystudio.com/)。它们通过自然语言处理技术从文本中提取信息和知识,同时聘请专业人士进行判读,保证知识的可靠性。IBM 和微软等公司依托其在人工智能及信息处理技术等方面的强大优势,研发了医学知识智能检索、查询和相关分析工具,代表性产品有 Watson 肿瘤治疗和临床应用系统、微软 Microsoft Health 系统[8]。IBM Watson 是一种数据分析软件,可以自动化分析、预测分析和可视化分析,但是需要数据预处理、统计概念理解和领域专业知识[8]。

以上产品或平台为生物医学研究和药物开发提供了高效广泛的解决方案,覆盖了从药物发现到新药申请,再到临床试验和临床辅助诊断等生物医学各个方面。总的来说,以基因为中心,描述基因-基因、基因-疾病、基因-位点关联的知识库越来越多,其对精准医学的巨大价值也得到了广泛的认同。

2.2 国内生物医学知识库建设

我国生物医学知识库建设也已经起步，主要是基于单一信息来源的医学知识库建设。以文献知识库为代表的医学知识库广泛服务于临床研究机构和临床医生，如以文献知识库为代表的中国生物医学知识库（中国医学科学院医学信息研究所）、中国疾病知识总库 CDD（军事科学院图书馆）、中国医院知识总库 CHKD（中国知网）以及临床诊疗知识库（万方医学网）[9]。“中国医院知识总库”和“中国疾病知识总库”是拥有多检索入口、分组和排序、库间引文链接、知识网络等功能的全文知识库[10]。“中国疾病知识总库”，不仅面向临床医药学专业人员，而且兼顾普通大众；万方医学网是完善的临床诊断的知识库，提供以疾病、症状、检查、药品、指南和病例报告为基础整合的知识点，方便医生查找相关知识及病例，辅助医生临床诊断[10]。以上医学知识库为临床诊疗和研究提供的知识服务源于可直接利用的知识，但是没有进行知识识别和知识推理。近年来随着知识库构建和知识图谱技术的发展，各种基于临床病历和专科专病治疗的专题知识库的完善，以及专业医学知识库对临床辅助诊断应用的重要性凸显，具备知识推理和发现的医学知识库成为了研究热点。

此外，还有一些自主构建的知识库，如复旦大学和上海生物信息技术研究中心完成的“面向基层医疗基于循证医学的知识库系统”、中国人民解放军军事医学研究院构建的肝癌知识库、浙江大学开发的个性化合理用药系统和智能诊疗协议推荐系统等。但是这些知识库还需要在数据完备、标准共享等方面进行完善，从而与国际接轨。值得注意的是，目前尚无针对中国人群的生物医学知识库，也没有与 GeneGo、IPA 等比肩的应用平台，掣肘我国精准医学发展。为了打破国外生物医学知识库的垄断，更好地为中国精准医学提供支撑，以复旦大学为首的团队拟在国家重点研发项目中利用标准的语义网络，获取完整全面的精准医学信息，并运用先进的知识发现技术，构建统一的、开放共享的、高效准确的精准医学知识库，从而服务于中国精准医学。

3　疾病研究精准医学知识库

3.1　构建目标

复旦大学承担的隶属于国家重点研发项目的精准医学项目“疾病研究精准医学知识库构建”主要针对精准医学知识数据量庞大、数据类型复杂、资源分布不均衡、利用程度低下等问题，立足我国多层次的精准医学知识库体系和安全稳定可操作的生物医学大数据共享平台的建设需求。其主要目标是面向恶性肿瘤、心脑血管疾病等全疾病谱，整合生物医学本体和多类型医学文本资源，融合多层次生物信息数据，加工和分析海量异构异源生物医学本体和生物信息资源，分析生物通路和网络特征，构建规范化、结构化、自动更新和多维自动化与人工审编的精准医学知识库体系，形成对精准医学研究和临床应用的关键支撑。开发的精准医学知识库体系填补了国内空白，打破了国际垄断，将为针对健康和疾病人群的精准医学研究和临床应用提供多层次支撑。

3.2　构建原则

集成深度索引、相关性挖掘、重要性标注、新颖度分析等挖掘工具，集大规模文献整合分析与知识发现于一体，实现精准医学知识的抽提、注释、聚类、关联及分析，开展基于数据关联和可视化的精准医学知识利用技术研究，实现基于生物医学语义和本体的全文检索、文本识别、关键词分析等功能。从海量的组学数据和临床数据出发，对生物医学知识进行跨库融合，并通过大数据网络特征分析技术、模型特征提取技术等生物信息学手段与工具的开发，利用知识图谱构建与扩展技术的应用，构建“基因-通路-疾病-症状-诊疗-药物”的精准医学知识图谱，形成面向精准医学的疾病相关生命组学知识库体系。进一步开发知识网络和知识图谱的多维自动注释流程，建立协同审编平台，形成精准医学知识库。最后，开发可交互、定制、扩展、自动更新的工作流技术体系，在“精准医学大数据平台”上实现精准医学知识库的检索、展示、管理与共享，以及面向科研与临床不同需求的知识库应用。汇集大规模文本挖掘、疾病相关生命组学、第三方知识库等证据源形成的知识，构建精准医学知识整合模型，实现精准医学知识的自动化注释，并研究开放式的精准医学知识人工审编技术，建成多证据源整合的疾病相关精准医学知识库体系，开发检索与展示功能，搭建精准医学知识库管理与共享平台。

3.3　总体框架

构建精准医学本体和语义网络，建立精准医学文本知识网络。通过跨库融合、大数据网络特征分析、模型特征抽取等手段，整合多种生物信息数据，构建和扩展精准医学知识图谱。整合精准医学知识网络和知识图谱，建立面向文本和组学数据自动注释与融合的流程，基于多维度的证据进行人工审编，形成精准医学知识库；对接“精准医学大数据平台”，实现个性化检索、展示和自动更新，支撑面向精准医学的知识服务。具体研究内容分为以下 5 部分。

3.3.1 精准医学本体和语义网络构建

借鉴 ICD-10、MeSH、UMLS 等生物医学本体，建成涵盖组学、疾病、症状、药物等科技词表和本体的规范精准医学语义关系，形成标准化、结构化的精准医学知识模型。设计并建立精准医学知识组织框架，集多来源医学知识组织系统为一体的建设方案和技术路线，开展精准医学领域术语采集，实施精准医学领域术语遴选与清洗，对遴选的精准医学领域术语进行评价并进行结构转化。

开发精准医学本体协同加工系统，实现多来源的异构异型词表导入与关联、词表和本体的可视化与交互式编辑，研发复杂本体的概念归并、同义关系相似度计算、不同概念间语义相关度计算和语义推理工具，为构建并维护疾病相关组学本体和语义网络提供有效工具。开发精准医学本体和语义网络共享服务接口，形成标准化、结构化的精准医学本体元数据集，为知识库建设提供灵活调用和模块式集成方式。

3.3.2 精准医学文本挖掘与知识网络构建

开展国际公开文献、专利、临床试验、药品监管等海量多源异构文本资源的采集、加工和规范化研究，建立精准医学文本资源数据库，实现全文检索、关键词分析和自动更新。根据系统构建的精准医学本体元数据集，定义医学文本的实体识别与关联抽取标注规范，开展工具标注与人工修正研究，构建更大规模、更高质量的精准医学文本训练语料库。基于融合词性信息、生物实体识别文本表示，利用海量的未标注生物医学文本训练词向量，自动学习更抽象更有效的特征，构建高性能的实体识别模型。

利用深度学习的方法和已建成的精准医学实体关联语料库，采取卷积神经网络进行实体语义关联抽取。利用相关性挖掘、高维聚类分析和关联网络构建技术，实现面向精准医学的大规模文献整合分析与知识发现，并应用于恶性肿瘤、心脑血管疾病等全疾病谱。

3.3.3 精准医学知识图谱的构建

通过收集、组织、整理与疾病发生、发展、治疗和预后相关的基因组学、转录组学、表观遗传组学、蛋白质组学和代谢组学等多组学数据的国际生物医学数据库和来源于大型临床机构的临床数据，利用跨库知识融合技术，初步构建涵盖“基因-通路-疾病-症状-诊疗-药物”关联关系的精准医学知识图谱。开展基于知识图谱的自动化补全技术，填补知识关联缺失值，完成精准医学知识图谱的第一层扩展。针对生物医学大数据形成的网络或模型，开发生物信息学算法，利用网络特征分析、模型特征提取，预测生物医学大数据的关联性，完成精准医学知识图谱的第二层扩展。开展数亿级别的海量知识图谱查询和检索技术研究，构建基于生物信息学的精准医学知识图谱，发展生物信息学通路和注释知识体系，对接搭建的大型开源生物通路数据库和系统生物学数据分析挖掘平台。

3.3.4 精准医学知识自动化注释与人工审编

开发面向精准医学知识库构建的基础数据接口与 ETL 工具集，整合大规模文本挖掘、疾病相关生命组学、第三方知识库等来源的精准医学知识，研究基因、蛋白、遗传变异、疾病、表型、药物等维度的实体异构知识的数据整合模型，研究不同证据源的精准医学知识热度和质量评价算法。开发精准医学知识自动化注释软件，选择代表性的疾病组织专家进行知识的人工审编，构建小规模、高质量的精准医学知识库。

开发开放式的精准医学社区平台，构建面向知识发现的全疾病谱精准医学知识库，建立多证据源融合的精准医学知识数据索引，提供基因、蛋白、遗传变异、疾病、表型和药物等不同维度的高效检索和筛选服务，为用户提供直观友好和易懂可读的知识展示。研发基于同质和异质的知识网络的知识发现方法，支持文本挖掘、生命组学和第三方知识的开放式注释等证据源的回溯。

3.3.5 精准医学知识库管理与共享平台研发

对接“精准医学大数据平台”，研发知识库信息资源管理系统，实现精准医学知识库信息资源的管理和分类展示，为各种知识库应用提供访问入口。开发可交互、定制、扩展、更新的工作流技术服务体系，整合项目产出的知识库工具，实现科研数据的处理、分析以及对接知识库服务。以基因、蛋白质为核心，研发基于精准医学知识库的通路和网络的结果展示、重要成份标注、功能注释和精细化作图等技术。面向医学基础研究和临床实践需求，开发个性化的知识推送系统和开放性的知识库应用接口（API）服务，覆盖典型的精准医学知识查询。研究精准医学知识临床转化关键技术，在临床机构建立典型应用示范。

3.4　关键技术

3.4.1　构建复杂生物医学本体集成与标准化的精准医学知识模型

利用 Protégé 等本体构建工具框架和本体映射技术，集成复杂生物医学本体对现有的生物医学领域本体进行规范化，实现多来源词表的统一存储与关联。面向本体中多类概念以及复杂语义关系，采用词汇级、短语级精准医学词汇的映射算法，实现疾病、基因、蛋白质、药物、环境、通路等术语的概念归并。

标准化的精准医学知识模型构建则是通过开发精准医学概念、属性、语义关系和唯一标识符控制工具，系统构建并维护疾病相关组学本体和语义网络。借鉴数据交换、知识表示存储的国内外行业标准，形成标准化、结构化的精准医学本体元数据集，实现精准医学本体和语义网络共享和接口调用。

3.4.2　构建精准医学文本实体识别和语义关联抽取模型

构建生物医学文本实体识别模型。针对传统词袋模型存在的维度高、数据稀疏、忽略词序信息等问题，提出基于融合词性信息、生物实体识别文本表示，利用海量的未标注生物医学文本训练词向量，学习词语间丰富的内部关联；利用结合状态转移概率的双向长短期记忆 LSTM 神经网络，自动学习更抽象更有效的特征，构建高性能的生物医学文本实体识别模型。

构建生物医学文本语义关联抽取模型。针对现有词向量大都基于线性词序的上下文关系，忽略实体关系抽取中重要的句法信息的问题，提出基于句法词向量的文本表示方法，将其输入到卷积神经网络中，通过深度学习模型学习有效的特征，提升实体关系抽取的性能。

3.4.3　构建基于生物信息学的精准医学知识图谱

精准医学知识图谱的自动化补全研究。依据精准医学知识图谱的子结构特征与相关生物医学资料的关系，利用人工和机器学习两种方法，基于对应模板提取相应知识点，自动填补知识关联缺失值，弥补人工构建知识图谱的局限性。

基于生物信息学的生物数据关联挖掘研究。基于分类、回归分析、时间序列分析、聚类、关联分析和序列分析等生物数据挖掘方法，寻找生物组学数据与临床诊断、疾病分型、预后分析、药物开发等医学研究与实践的关联关系，并将以上关系以特殊标记属性值的形式补充在精准医学知识图谱中。

构建基于生物信息学的精准医学知识图谱。针对生物医学概念识别的异构特征，如类别相似度、语义相似度和图结构相似度，归并分散的知识，增强对生物医学概念的多层面理解。通过知识的跨库融合实现从单纯的知识库整合到知识图谱网络构建的跳跃。

3.4.4　构建精准医学知识自动化注释与人工审编

多证据源的知识整合技术。精准医学知识来源多样，既有来自采用大规模自动挖掘得到的基于文本的知识和基于疾病相关生命组学数据挖掘得到的知识，也有来自第三方的经过审编的知识。不同来源的知识可能存在冲突，需要对证据源进行有效的整合。

建立开放式的精准医学知识审编社区。精准医学知识可以按照基因、蛋白、疾病等实体来组织，也可以按照文本资源来组织。2 个角度的组织方式都支持对知识的评价和纠错机制，引导外部志愿者改进知识质量。

精准医学知识库检索与展示技术。采用 MongoDB 的 NoSQL 技术，将不同证据源的知识模型优化为简单的以键值对为核心的分布式结构，并引入高效的非结构化文本资源的搜索引擎 Solr，提高数据检索性能和扩充性能，满足知识库检索和展示的

需求。

3.4.5 构建精准医学知识库的管理与共享系统

精准医学知识库管理共享平台基础架构。基于开源 Galaxy 框架进行二次开发构建工作流体系，通过自动化数据处理和人机交互数据处理方式实现数据库更新。

精准医学知识分析和精细作图体系构建。借鉴 Reactome 等在知识分析和精细作图体系方面的优势，对标 GeneGo/IPA，搭建对通路信息进行分析和可视化的工作平台。

精准医学知识库推送和应用接口（API）的开发和应用示范。针对典型精准医学知识需求，利用 WebService 开发以 JSON/XML 等标准发布数据的应用接口（API），并基于此建立个性化、智能化的知识订阅和自动推送机制，支撑精准医学临床决策支持并开展精准医学知识库临床评测。

3.5 实现功能

本项目预期建成面向疾病研究的精准医学知识库体系。该体系具有标准规范、开放共享、用户友好、动态更新等特点，并可作为国家标准来促进疾病精准医学研究成果的知识管理。在此过程中，将取得“一个标准”“一个库”“一个平台”3 大成果。

3.5.1 精准医学本体和语义表示标准

参照国际上通用的 ICD-10 分类法及 UMLS、MeSH 等生物医学主题词表和本体，建成涵盖组学、疾病、症状、药物等的科技词表和本体，构建精准医学语义网络和知识模型，形成标准规范、系统全面的精准医学本体和语义网络标准。

3.5.2 精准医学知识库

根据系统构建的本体和语义网络，针对海量的多源异构文本和生物信息数据，利用自然语言处理、深层索引、相关性挖掘等技术进行数据整合、关联抽取，形成“精准医学文本知识网络”和“基于生物信息学的精准医学知识图谱”。通过自动注释和审编，并结合重要性标注和新颖性分析，实现文本与组学数据的融合和知识发现，最终形成涵盖多证据源的，面向恶性肿瘤、代谢系统疾病、呼吸系统疾病、心脑血管疾病、免疫性疾病、神经精神类疾病等疾病的，覆盖科学研究和临床应用等需求的，可实现证据分级与回溯功能的精准医学知识库。

3.5.3 精准医学知识库管理与共享平台

该平台将以网站的形式呈现，支持面向精准医学知识发现的检索、展示和分析功能（对标 GeneGO 和 IPA），提供不少于 20 个用于知识抽提、文本挖掘、信息展示和精细作图等的软件工作流技术服务体系，支持面向用户定制的知识推送，提供面向科研和临床等多种场景的应用接口。本平台将部署在“精准医学大数据平台”的服务器和云端，稳定支持 1000 人以上并发使用，且具有自动更新功能。本平台将通过“精准医学大数据平台”向本专项所有项目推广、开放、共享。

4 PMapp 精准医学知识库介绍

精准医学知识库（Precision Medicine knowledge-base application，PMapp）以知识地图的形式全面整合了 45 个数据库（图 1），其主要框架包括基因及其产物、生物信号通路和分子网络、致病变异以及药物 4 个部分（图 2）。

在第一部分中，PMapp 收存了 20 656 个编码人类基因和 38 943 个非编码人类基因，178 562 个 RNA 以及 111 716 个蛋白质。这些基因及其代谢产物构成了 PMapp 实体存储库的主要部分。至 2017 年底，第二部分已经集成 21 个信号通路/网络数据库，其中包含 13 个主要相互作用类别、22 种生物效应、28 种修饰和 1 个实验注释。

整个常规通路和网络总共涵盖 31 264 个生物实体（节点）和 1 804 000 个相机作用（边），包含 13 种不同的作用关系（表 1）。致病变异方面，PMapp 收集了 5 738 719 种致病变异，源自 18 022 个基因，对应10 725种疾病。

除此之外，9 746 种药物和其对应的 78 664 个靶标的信息也被收录在 PMapp 中，包含561 180个药物-药物，1 191 个药物-食物，5 118个药物-酶，以及1 839种药物-转运体等相互作用。PMapp 在本体方面实施面向精准医学的重大疾病本体体系结构、知识表示模型和精准医学术语库的构建。精准医学词汇规模达到 300 万，整合了 57 部生物医学领域词表、术语 3 879 621 个、概念1 052 512个。完成精准医学本体的语义概念映射，精准医学本体覆盖 2 个重大疾病领域，本体之间的语义关系包含疾病-基因-药物。完成精准医学本体构建，精准医学本体

包括类 57 746 个、语义关系 92 个，涵盖人类表型、疾病、化学物质与药物、细胞机制、分子机制、遗传机制 6 大医学领域，完成2 ~3个重大疾病精准医学本体建设。

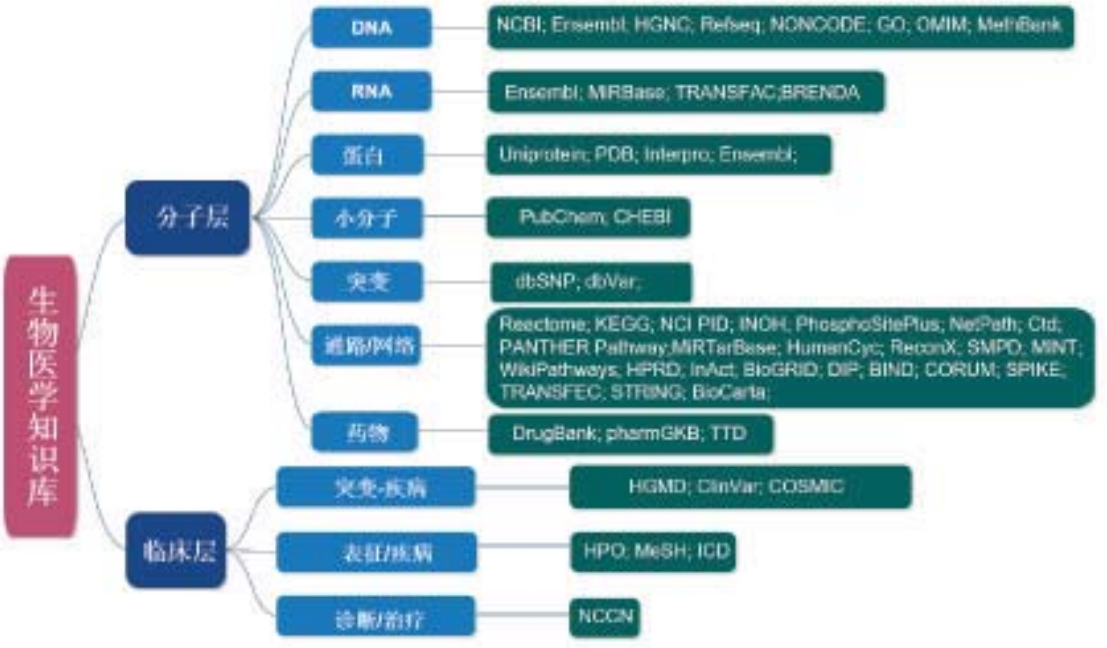

图 1　PMapp 完成 54 个不同领域的数据库数据收集和整理

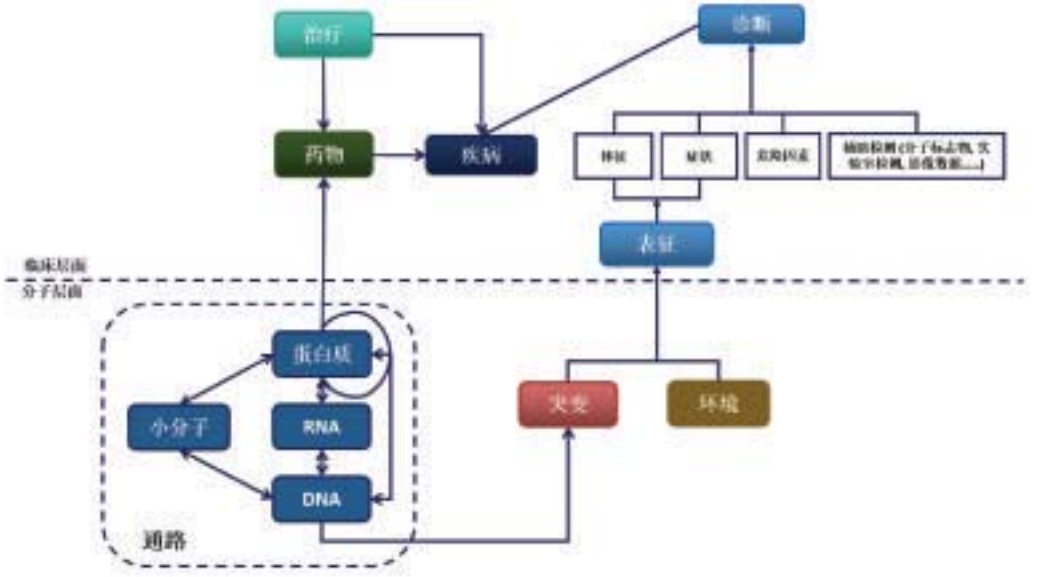

图 2　精准医学知识库的主要框架

表 1 知识库所收录的相互作用类型

Pathway Type	Edge Type	Directionality	Edge No.
Signaling Pathway	SR: Signaling Regulation	Directed	172 765
	ER: Expression Regulation	Directed	122 786
	CAI: Complex Assembly Interaction	Undirected	177 227
	TR: Transport Regulation	Directed	7 296
	TRc: Transport Regulation chemical	Directed	3 285
	ca: x chemical affects P	Directed	469 519
	RNAi: RNA interference	Directed	317 556
TechPPI	TechPPI: Technical Protein-Protein Interaction	Undirected	316 437
Metabolic Pathway	sp: metabolic reaction s $\xrightarrow{E}$ p $\xrightarrow{F}$	Directed	14 428
	sE: metabolic reaction s $\xrightarrow{E}$ p $\xrightarrow{F}$	Directed	22 480
	Ep: metabolic reaction s $\xrightarrow{E}$ p $\xrightarrow{F}$	Directed	21 334
	EE: metabolic reaction s $\xrightarrow{E}$ p $\xrightarrow{F}$	Directed	154 975
	rw: x reacts with y	Undirected	3 912

目前 PMapp 网站集成了项目产出的本体和术语资源以及数据库资源，搭建了精准医学知识库检索网站，可以检索基因、药物等主要精准医学概念。在基因展示页面中，对基因基本信息、GO 注释信息、蛋白质结构、蛋白质相互作用、关联药物、关联疾病等分别做了显示。因此，PMapp 是面向科学研究的，可以进行通路/网络精细做图、通路分析和智能检索的系统。

此外，PMapp 将无缝衔接本体富集分析等已有的分析工具，对标 IPA 基本的工作流分析体系，可以实现对组学数据的差异基因分析，并给出火山图等图表；也会在系统中整合基因本体、通路本体、疾病本体、Mesh 本体 以及本项目产出的综合本体等资源数据，对数据进行本体注释及富集分析，并提供可呈现上下层级关系的本体树图形及表格多样化的结果展示。

针对用户关注或感兴趣的组学数据或基因列表，本系统可为用户提供一系列的注释信息，比如本体、分子网络等。将要实现的本体表格和 DAG 树状结构的展示以及分析功能，对标 IPA/GeneGO 的精准医学知识库通路分析展示系统的通路展示和分析。通过搜索页面找到需要的通路，然后通过网络展示页面显示通路分子成份之间的关系，也可以对网络进行编辑和聚类分析等。

在这个过程中，项目团队首先针对精准医学知识库如 PharmGKB 提供了一种新的知识查询服务。目前最大的问题是现有的标准化知识访问协议通常都是基于标准术语系统和编码来检索的，而目前临床数据通常不太具备这样的数据质量。后续需要改进中文术语的标准化编码查询服务，以满足知识精准查询的需求，实现 PMapp 双重角色，即为针对健康和疾病人群的精准医学研究和临床应用提供多层次支撑。

5 精准医学知识库展望

实现精准医学的核心，是结合多维度的临床、影像和多组学数据，利用深度学习、自然语言处理、多组学整合分析等方法，研发面向疾病风险预测、早期诊断、精准治疗、疗效评估和预后监控的精准医疗临床决策支持系统，为精准医学临床应用转化提供有效途径，从而实现疾病精准预防、精准检测和精准诊疗的目标。

通过知识库的构建和完善，完成多学科协作、贯通诊疗全过程的精准医疗临床决策支持系统。该系统具备多模态信息融合、影像组学联合分析、深度学习决策模型集成、多尺度决策硬件加速和多环节全景式分析等能力，可实现覆盖多学科及完整就医闭环流程的精准医疗辅助决策，依托大型综合性医院验证推广，将明显提高恶性肿瘤、代谢系统疾病、呼吸系统疾病、心脑血管疾病、免疫性疾病、神经精神类疾病和罕见疾病等重大疾病的医疗水平和防治效益。

【参考文献】

[1] Bourne PE, Lorsch JR, Green ED. Perspective: Sustaining the big-data ecosystem[J]. Nature, 2015, 527(7576): S16-17.

[2] Perez-Riverol Y, Alpi E, Wang R, *et al*. Making proteomics data accessible and reusable: current state of proteomics databases and repositories[J]. Proteomics, 2015, 15(5-6): 930-949.

[3] Berger ML, Lipset C, Gutteridge A, *et al*. Optimizing the leveraging of real-world data to improve the development and use of medicines[J]. Value Health, 2015, 18(1): 127-130.

[4] Argyropulo-Palmer M, Jenkins A, Theti DS, *et al*. Sunitinib in Metastatic Renal Cell Carcinoma: A Systematic Review of UK Real World Data[J]. Front Oncol, 2015, 5: 195.

[5] Landrum MJ, Lee JM, Benson M, *et al*. ClinVar: public archive of interpretations of clinically relevant variants[J]. Nucleic Acids Res, 2016, 44(D1): D862-868.

[6] Orchard S, Ammari M, Aranda B, *et al*. The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases[J]. Nucleic Acids Research, 2014, 42(D1): D358-D363.

[7] Joshi-Tope G, Vastrik I, Gopinath GR, *et al*. The genome knowledgebase: A resource for biologists and bioinformaticists[J]. Cold Spring Harbor Symposia on Quantitative Biology, 2003, 68: 237-243.

[8] Hoyt RE, Snider D, Thompson C, *et al*. IBM Watson Analytics: Automating Visualization, Descriptive, and Predictive Statistics[J]. JMIR Public Health Surveill, 2016, 2(2): e157.

[9] 蒋立辉, 王　伟. 医学知识库与医学知识的获取[J]. 医学信息(上旬刊), 2006, 19(9): 1500-1502.

[10] 马　利, 崔志伟, 毛树松. 我国医学知识库应用现状研究[J]. 医学信息学杂志, 2013, 34(11): 55-59.

[收稿日期: 2018-03-20]

[本文编辑: 施沅坤]